

EFFICIENT AND RELIABLE VIRTUAL VOLUME MAPPING

George R. Cameron

Hueichian Huang

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application Nos. 60/408,735 and 60/409,311, both filed September 6, 2002, which are incorporated herein by this reference.

TECHNICAL FIELD OF THE INVENTION

[0002] The present invention relates to data storage systems, and in particular, to efficient and reliable virtual volume mapping.

BACKGROUND

[0003] Network storage systems may be composed of multiple interconnected storage devices connected to one or more servers to increase performance in terms of storage capacity, reliability, scalability, and availability. System storage performance can be enhanced by using system management operations including management of data communication and management of data placement. Data system management has multiple aspects including storage device mapping, methods of data storage and recovery, data integrity, backup operations, and storage element utilization.

[0004] Network storage systems can store large amounts of data at least partly on the basis of inherent and simple scalability. The volume capacity can be simply increased by adding physical storage devices. However, the mere addition of hardware resources does not create the most efficient and useful storage system. Storage management techniques may be used, for example, to operationally include added storage capacity into a storage system, organize the storage for usage, distribute data in the storage, replicate data as needed, and assure data security.

[0005] Virtualization of memory is one aspect of storage management that permits allocation of storage resources according to dynamically changing requirements. Virtualization may be executed in a storage management software layer located between physical device drivers and a file system to create a virtual volume description that is not constrained to specific physical

storage devices. The virtualization software describes the physical storage as a pool of resources that can be dynamically managed.

[0006] Various techniques and approaches have been attempted for virtualization. But these have suffered from various disadvantages. What is needed are virtualization techniques that improve, for example, data accessing performance, promotes space-efficient data storage, and facilitates data sharing.

SUMMARY

[0007] Disadvantages and problems associated with previously developed systems and methods for data storage have been substantially reduced or eliminated with various embodiments of the present invention.

[0008] In one embodiment, a snapshot tree structure includes a base volume storing a current user data, a first snapshot descending from the base volume, and a second snapshot descending from the first snapshot, wherein the second snapshot was created prior to the first snapshot. The first snapshot includes a first table corresponding to first portions of block addresses, a second table corresponding to second portions of the block addresses, and a third table to third portions of the block addresses. The first table includes an entry with a pointer to the second table. The second table includes an entry with a pointer to the third table. The third table includes an entry with a pointer to a data of the base volume before the data is modified in the base volume.

[0009] Important technical advantages of the present invention are readily apparent to one skilled in the art from the following figures, descriptions, and claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] Embodiments of the invention may best be understood by referring to the following description and accompanying drawings, in which:

[0011] FIGURE 1 illustrates a scalable cluster data handling system, which can be an exemplary environment within which embodiments of the present invention may operate.

[0012] FIGURE 2 illustrates a node for use in scalable cluster data handling system shown in FIGURE 1.

[0013] FIGURE 3 is a block diagram of a scalable cluster data handling software architecture.

[0014] FIGURE 4 is a schematic block diagram that illustrates an example of the use of a virtual volume region table for handling data in a data storage management system, according to an embodiment of the present invention.

[0015] FIGURE 5 illustrates the creation of a logical disk from the storage space of multiple physical disks.

[0016] FIGURE 6 illustrates the access to the virtual volumes of multiple nodes by a host device through the virtual volume region tables on several nodes.

[0017] FIGURE 7 illustrates a number of data structures for a snapshot technique that may be created as data is written or modified in a base virtual volume.

[0018] FIGURE 8A illustrates one view of storage volumes for the data structures of FIGURE 7.

[0019] FIGURE 8B illustrates another view of storage volumes for the data structures of FIGURE 7.

[0020] FIGURE 9 is a flowchart for an exemplary method for a multiple level mapping for a virtual volume, according to an embodiment of the present invention.

[0021] FIGURE 10 is a flowchart for an exemplary method for a delayed copy-on write (COW) operation for a virtual volume, according to an embodiment of the present invention.

DETAILED DESCRIPTION

[0022] The preferred embodiments for the present invention and their advantages are best understood by referring to FIGURES 1 through 10 of the drawings. Like numerals are used for like and corresponding parts of the various drawings.

[0023] Turning first to the nomenclature of the specification, the detailed description which follows is represented largely in terms of processes and symbolic representations of operations performed by conventional computer components, such as a local or remote central processing unit (CPU), processor, server, or other suitable processing device associated with a general purpose or specialized computer system, memory storage devices for the processing device, and

connected local or remote pixel-oriented display devices. These operations may include the manipulation of data bits by the processing device and the maintenance of these bits within data structures resident in one or more of the memory storage devices. Such data structures impose a physical organization upon the collection of data bits stored within computer memory and represent specific electrical or magnetic elements. These symbolic representations are the means used by those skilled in the art of computer programming and computer construction to most effectively convey teachings and discoveries to others skilled in the art.

[0024] For purposes of this discussion, a process, method, routine, or sub-routine is generally considered to be a sequence of computer-executed steps leading to a desired result. These steps generally require manipulations of physical quantities. Usually, although not necessarily, these quantities take the form of electrical, magnetic, or optical signals capable of being stored, transferred, combined, compared, or otherwise manipulated. It is conventional for those skilled in the art to refer to these signals as bits, values, elements, symbols, characters, text, terms, numbers, records, files, or the like. It should be kept in mind, however, that these and some other terms should be associated with appropriate physical quantities for computer operations, and that these terms are merely conventional labels applied to physical quantities that exist within and during operation of the computer.

[0025] It should also be understood that manipulations within the computer system are often referred to in terms such as adding, comparing, moving, searching, or the like, which are often associated with manual operations performed by a human operator. It must be understood that no involvement of the human operator may be necessary, or even desirable, in the present invention. Some of the operations described herein are machine operations performed in conjunction with the human operator or user that interacts with the computer or system.

[0026] In addition, it should be understood that the programs, processes, methods, and the like, described herein are but an exemplary implementation of the present invention and are not related, or limited, to any particular computer, system, apparatus, or computer language. Rather, various types of general purpose computing machines or devices may be used with programs constructed in accordance with the teachings described herein. Similarly, it may prove advantageous to construct a specialized apparatus to perform one or more of the method steps described herein by way of dedicated computer systems with hard-wired logic or programs stored in non-volatile memory, such as read-only memory (ROM).

Overview

[0027] According to embodiments of the present invention, systems and methods are provided for virtual volume creation, configuring, mapping, and management in a data storage system. This can, among other things, facilitate the addition of storage capacity into the data storage system, organize storage space for usage, distribute data in the storage space, replicate data, and maintain data integrity. The systems and methods may perform or support data sharing operations for backup protection, failure recovery, and data mirroring performance.

[0028] According to one aspect of the present invention, a data storage system is capable of execution in one or more nodes that couple one or more host systems to a plurality of data storage devices. The data storage system may include volume manager that may create, control, or manage a logical disk abstraction of one or more of the plurality of data storage devices. The volume manager may utilize or be used in conjunction with one or more virtual volume mapping tables. The virtual volume mapping tables comprise a plurality of entries that correspond to storage regions in the plurality of data storage devices. The volume manager may create mappings from the virtual volume mapping tables to the logical disk abstraction.

[0029] In another aspect, the data storage devices of a data storage system may be of various sizes and speeds. The volume manager may be capable of creating mappings in the virtual volume mapping tables according to data storage device size and speed. The data storage devices may be divided into “chunklets,” having various performance specifications. The physical chunklets may be grouped into the logical disk abstraction according to the performance specifications.

[0030] In accordance with another aspect of the invention, the virtual volume can provide a logical disk abstraction that represents different physical disks on different physical disk sleds in the data storage system. Failure or removal of any physical disk or disk sled does not make data unavailable. In accordance with a further aspect of the invention, the logical disk abstraction of the virtual volume may represent the physical regions of a plurality of data storage devices with different interface controllers and interfaced to different nodes of the data storage system.

Exemplary Environment

[0031] FIGURE 1 illustrates a scalable cluster data handling system 10, which can be an exemplary environment within which embodiments of the present invention may operate. The scalable cluster data handling system 10 is an architecture suitable for communication-intensive,

highly available data storage, processing, and/or routing. The architecture is useful for many applications to provide high performance, scalable, flexible, and cost-effective storage arrays.

[0032] Scalable cluster data handling system 10 can be incorporated or used in a data storage system to provide mass storage for data and information routed, generated, manipulated, processed, or otherwise operated upon, by various host devices 18. These host devices 18 can include various processing devices, such as, for example, server clusters, personal computers, mainframes, and server towers. Host devices 18 may also include various peripheral devices, such as, for example, printers, a modems, and routers. Each of these host devices 18 is connected to scalable cluster data handling system 10. As used herein, the terms "connected" or "coupled" mean any connection or coupling, either direct or indirect, between two or more elements; such connection or coupling can be physical or logical. The data storage system (in which scalable cluster data handling system 10 may be incorporated) also includes a number of storage devices 20. These storage devices 20 can be implemented with any suitable mass storage resource, such as tape or disk storage. In one embodiment, the storage devices 20 may be one or more JBOD (Just a Bunch of Disks) facilities comprising a plurality of disk drives. The disk drives can be mounted in a rack-mountable storage shelf having one or more hot-pluggable disk drive sleds. Each sled may accommodate four disk drives on a pair of fibre channel (FC) connections. The sleds can be configured in one of two possible ways: (1) all sleds on the same redundant FC connections, or (2) half of the sleds on one set of redundant FC connections and the other half of the sleds on another set of redundant FC connections. Scalable cluster data handling system 10 allows the host devices 18 to store and retrieve information from the storage devices 20.

[0033] As depicted, the scalable cluster data handling system 10 includes a plurality of interconnected nodes 12. In the illustrative example, eight nodes 12 are provided, with each node 12 connected to every other node 12 by a respective high-speed link 16. Each node 12 generally functions as a point of interface/access for one or more host devices 18 and storage devices 20. In an illustrative example of the scalable cluster data handling system 10, a node 12 can be a modular computer component that has one or more PCI bus slots or other interfaces for connections to storage devices 20 and host devices 18. For this purpose, in one embodiment, each node 12 may include one or more peripheral component interconnect (PCI) slots, each of which supports a respective connection 14. Each connection 14 can connect a host device 18 or a storage device 20. Connections can be small computer system interface (SCSI), fibre channel (FC), fibre channel arbitrated loop (FCAL), Ethernet, Infiniband, or any other suitable

connection. A node 12 performs software processes (procedures, methods, and routines) under control of independently executing operating systems.

[0034] In the illustrative example of FIGURE 1, a host device 18 (i.e., Host 1) is in communication with a plurality of nodes 12 (i.e., Node 0 and Node 1). These nodes 12 control access to a plurality of storage devices 20 (e.g., physical disks) that are separated into multiple storage regions. A virtual volume mapping or table (described herein) at each node 12 comprises pointers that are configured to designate the location of data on the storage devices 20. The host device 18 accesses all of the storage devices of the region table in the manner of accessing a single large, fast, and reliable virtual disk with multiple redundant paths. Host device 18 can write to a particular storage region on the storage devices via one communication path and read back data on a different path. The virtual volume region table is used to track the stored data so that the most recent and correct data copy can be accessed by the host device 18 from a proper node 12.

[0035] FIGURE 2 illustrates a node 12 for use in scalable cluster data handling system 10 shown in FIGURE 1. Node 12 supports connections 14 for connecting host devices 18 and storage devices 20, and communication paths 16 for communicating with other nodes 12. As depicted, node 12 can be implemented with a computer-memory complex 30, a node controller 32, and a cluster memory 34. Computer-memory complex 30 can be a computer system which includes one or more central processing units (CPUs) and associated memory running an independent copy of an operating system. Computer-memory complex 30 functions to support, control, or otherwise manage one or more suitable buses through which data and information can be transferred via connections 14. Node controller 32 and cluster memory 34 are distinct and separate from computer-memory complex 30. Node controller 32 may cooperate with computer-memory complex 30 but, to some degree, operates independently of the same. Cluster memory 34 is coupled to node controller 32 and generally functions to cache data and information being transferred through node 12. In the scalable cluster data handling system 10, cluster memory is used as a cache for disk reads and writes. Areas of cluster memory that are used for write caches are replicated on other nodes.

[0036] Additional details regarding the scalable cluster data handling system and its nodes are provided in co-pending U.S. Patent Application Serial No. 09/633,088, entitled "Data Storage System," (Attorney Docket No. M-8494) filed on August 4, 2000; U.S. Patent Application Serial No. 09/883,681, entitled "Node Controller For A Data Storage System" (Attorney Docket No. M-8496) filed on June 18, 2001; and U.S. Patent Application No.

entitled “Time And Space Efficient Technique for Creating Virtual Volume Copies” (Attorney Docket No. 3PD-M-8497-US), filed concurrently. These applications are assigned to the same Assignee as the present application and are hereby incorporated by reference in their entireties.

[0037] FIGURE 3 is a block diagram of a scalable cluster data handling software architecture 1000. This software architecture 1000 may be used in scalable cluster data handling system 10 of FIGURE 1. The scalable cluster data handling software architecture 1000 may be implemented on one or more nodes 12 and is configured to supply reliable, high-performance storage services for transferring data between host devices 18 (e.g., processors) and storage devices 20 (e.g., physical disk drives). The storage services can be either the abstraction of raw disks via Small Computer Serial Interface (SCSI) commands, for example over Fibre channel or parallel SCSI, or higher level access protocols or network data services such as NFS, CIFS/SMB or HTTP.

[0038] The scalable cluster data handling software architecture 1000 complements primary data storage functions with additional storage support functions such as server-less backup, remote mirroring, and volume cloning. In addition to the storage and storage support functions, the scalable cluster data handling software architecture 1000 supplies administration and management tools that automate tuning and recovery from failure, and supply centralized system management and monitoring.

[0039] A host interface layer 1004 connects the host devices to the scalable cluster data handling software architecture 1000. The host interface 1004 can include Fibre Channel/SCSI (FC/SCSI) target drivers 1010 and network adapter drivers 1012. File systems 1024 communicate with the host interfaces 1004 via network data services 1014. The network data services 1014 can include TCP/IP or UDP/IP 1016 services, as well as NFS 1018, CIFS 1020, and HTTP 1022. NFS 1018, CIFS 1020, and HTTP 1022 can be used to access the file systems 1024.

[0040] Storage in the scalable cluster data handling software architecture 1000 also includes one or more virtual volumes 1026, logical disks 1028, and physical disk layers 1032. Associated with the virtual volumes 1026 are caches 1030. The physical disk layers 1032 include physical disk drivers 1034, which provide an interface for physical disk drives. Physical disks are logically divided into pieces called “chunklets” (described below in more detail). In an illustrative embodiment, chunklets are fixed-size, for example 256MB contiguous segments of

disk space. The logical disks 1028 are connected to the FC/SCSI target drivers 1010 as well as the file systems 1024. Logical disks 1028 comprise multiple chunklets organized into groups. A logical disk driver (not shown) controls operation so that the chunklets in a RAID group are arranged on different spindles and, if possible, on different Fibre channel strings. Some spindles may not be connected to a Fibre channel loop on the node. The disk caches 1030 can also be abstracted as logical disks 1028.

[0041] Virtual volumes 1026 are representations of data storage. The virtual volumes 1026 are an abstraction of the scalable cluster data handling software 1000 that are accessible directly by the hosts devices via the FC/SCSI target drivers 1010, or accessible internally by the file systems 1024. Virtual volumes 1026 provide high-performance by virtue of performing caching and optimized RAID level mappings in addition to basic, uncached fixed-RAID service supplied by logical disk abstractions. A virtual volume manager 1040 may be in communication with various components of the scalable cluster data handling software architecture 1000. Virtual volume manager 1040 general functions to configure, set-up, and otherwise manage virtual volumes 1026. Virtual volume manager 1040 may map blocks (or regions) of the virtual volumes 1026 onto blocks on logical disks 1028. The mapping can be used to cache selected blocks of a virtual volume, place selected regions of the virtual volume 1026 on higher performance RAID groups, and create point-in-time images (snapshots) or clones of data on virtual volumes 1026.

Virtual Volume Management

[0042] FIGURE 4 is a schematic block diagram that illustrates an example of the use of a virtual volume region table 104 for handling data in a data storage management system 100, according to an embodiment of the present invention. The data storage management system 100 controls data management operations, and can be implemented as part of, for example, scalable cluster data handling system 10. The data storage management system 100 can be implemented, at least in part, as software.

[0043] The virtual volume region table 104 is associated with a virtual volume 1026, which is a virtual representation of data storage. In a data storage system served by data storage management system 100, the virtual volume 1026 may represent the collective storage space of a number of hardware storage devices (e.g., physical disk drives 112). The virtual volume region table 104 includes entries 105 for a number of regions (e.g., REGION 0, REGION 1, REGION2, ..., REGION N) of the virtual volume that correspond to particular storage spaces in the storage devices. In particular, the regions map to one or more logical disks 106 that provide access to a

plurality of physical disks 112. The virtual volume region table 104 may be used to record and manage the ownership of regions in a storage structure, such as the virtual volumes 1026, by a one or more nodes 12. In one embodiment, each virtual volume 1026 in a network is associated with its own ownership table.

[0044] The virtual volume region table 104 comprises an entry for each region in a virtual volume 1026. Thus, if a virtual volume has 100 regions, then the table 104 has 100 entries. Each entry in the virtual volume region table 104 stores an indication (e.g., an address) of an owning node (or owner) assigned to the region and an indication (e.g., an address) of a backup node (or replicant) assigned to the region. The owner is the node allocated to track a region of virtual memory stored in the physical storage associated with that owner node. A replicant node functions as a backup to track a region of virtual memory.

[0045] In one embodiment, each entry 105 of the virtual volume region table 104 includes one or more elements that provide pointers 108 that point to logical disks 106. As depicted, these elements may include an owning node element 130, a backup node element 132, a logical disk element 134, and a region element 136 for specifying the owning node, the backup node (or replicant), a logical disk, and a region, respectively. A pointer 108 (e.g., LD.id.reg_number) points to a particular logical disk (e.g., LD.id) and a particular region (e.g., reg_number) on the logical disk 106. The virtual volume 104 may thus virtualize all storage on multiple physical disks 112. Present-technology physical disks 112 may have a size of about 1GB to about 144GB, so that the virtualization of many physical disks creates an enormous storage space. From a host device's perspective, the virtual volume 104 may be accessed and behave in the manner of a physical disk.

[0046] The virtual volume 1026 associated with region table 104 may have a total virtual volume size that is substantially equal to sum of the storage capacity of the hardware storage devices represented. The regions of the virtual volume 1026 (e.g., REGION 0, REGION 1, REGION2, ..., REGION N) may each correspond to particular storage space. Each region may be the same or different size. In one embodiment, the number of regions in the virtual volume 1026 is equal to the total virtual volume size (e.g., 1 Terabyte) divided by the region size (e.g., 16 Megabytes) divided by the number of nodes (e.g., 8 nodes). Each region of the virtual volume 1026 may be associated with one or more logical disks 106 which, in turn, may be associated with one or more physical disks 112.

[0047] The virtual volume region table can be accessed by a host device 18. Furthermore, the virtual volume region table 104 may be accessible, via a volume manager 102, to a user interface 120 or an operating system 122. The user interface can be a graphical user interface (GUI) or a command line interface (CLI). The operating system 122 is local to a scalable cluster data handling system 10 or a node contained therein. This allows files of the virtual volume to be exported/imported using standard network file systems (e.g., Sun Microsystem's Network File System (NFS) and Microsoft's Common Internet File Service (CIFS)) associated with the operating system 122, or as web pages using hypertext transfer protocol (HTTP).

[0048] The volume manager 102 creates, configures, and manages the virtual volume (also called a "virtual disk") that is associated with virtual volume region table 104. To accomplish this, the volume manager 102 may create, modify, and delete entries of the virtual volume region table 104. The volume manager 102 may operate in Fibre Channel, small computer serial interface (SCSI), or other suitable interface, bus, or communication standard environments. In one embodiment, each node 12 in a scalable cluster data handling system 10 has its own separate volume manager 102. In another embodiment, a plurality of these nodes share one or more volume managers 102. In an illustrative example, the volume manager 102 presents the virtual volume (for example, over Fibre Channel) to one or more hosts 120.

[0049] The virtual volume is more reliable than physical disks because the volume manager 102 may implement a redundancy scheme that activates redundant replacement storage in the event of disk failure. The virtual volume can be much larger than a single physical disk and have a size that can change dynamically through operations of the volume manager 102. Also, the virtual volume can be enlarged in a relatively seamless manner. The virtual volume provides improved access performance and much lower latencies in comparison to physical disks, if for example, the virtual volume is accessed with patterns that are amenable to caching. The virtual volume may have a much higher bandwidth than physical disks. The virtual volume may be accessed over multiple interfaces, such as multiple Fibre Channels or SCSI links. Multiple interfaces for the virtual volumes permits performance of the virtual volume to exceed that provided by a single channel and allows continued availability of volumes following failure of one of the links. The virtual volume may be cloned to create copies of the original volume. Since any region of the virtual volume can be mapped to essentially any logical disk 106, the logical disks 106 can be configured to achieve specific performance criteria, depending on characteristics of the data access operations to be performed. Data access characteristics include

occurrence frequency of the operations, volumes of data to be processed, sparseness or concentration of data accessed in an operation, and the like.

[0050] The virtual volume allows placement of data in a self-tuning system to avoid data hot-spots. Virtual storage can be controlled so that portions of the virtual volume that are used frequently are moved to locations on logical disks 106, and thus physical disks 112, that are more easily and rapidly accessed. Some portions of the logical disks 106 have read access characteristics and write access characteristics that are different. For example, the characteristics may vary depending on the position of the data on a physical disk spindle. Outer locations and inside locations of a physical disk have different transfer rate characteristics on the basis that more data can be stored on the outer regions of a circular disk and less data stored in the interior regions. The amount of searching time to locate a particular data element may be reduced for inner regions.

[0051] In an illustrative operation, a host device 12 addresses the virtual volume as a single memory via the virtual volume region table 104. The region table 104 may map a region (e.g., REGION 0) of the virtual volume onto one or more logical disks 106 for any storage location. The volume manager 102 uses the virtual volume 104 to translate a virtual volume address to a logical disk address, and then to a physical storage location on a physical disk 112 by indexing into a virtual volume region table 104.

[0052] In one embodiment, the volume manager 102 creates the mappings from the regions of the virtual volume 104 (e.g., REGION 0 through REGION N) to one or more logical disks 106 and/or cached block locations in cluster memory (of one or more nodes). This allows logical disks 106 to be directly accessible by host devices 120. Mapping allows the virtual volume 104 to extend through multiple logical disks 106. Virtual volume mapping also allows an extremely large number of blocks to be cached with cache blocks located in the cluster memory of any node.

[0053] Virtual volume mapping enables additional storage functionality including creation of a “virtual volume clone” at another node or at another cluster data handling system. A “virtual volume clone” may be a copy of a virtual volume’s mapping, and can be both read and written. In one embodiment, when a virtual volume clone is first created, the virtual volume clone only includes a copy of the original virtual volume’s mapping, which is a small record that is quickly created and consumes almost no additional storage space. Accordingly, data of the virtual volume clone is accessed indirectly from the original virtual volume. When data is

written to the original virtual volume or the virtual volume clone, new physical storage blocks are allocated for the virtual volume clone. The mapping is changed only when particular disk blocks are written in either the virtual volume clone or the original virtual volume. If only a small fraction of total virtual volume size is written, then the additional memory space used by the clone is small.

[0054] An alternative technique for virtual volume cloning creates clones by physically copying the entire volume, which consumes the same physical disk space in the clone as is used in the original volume. Another alternative technique for virtual volume cloning utilizes read-only copying of a file system, not copying of physical storage. The read-only copies are adequate for some purposes such as backups, but read-write copies are required for purposes such as application testing on actual data.

[0055] The data storage management system 100 may be managed by backup software executed on the nodes. In an illustrative embodiment, the nodes run a general-purpose operating system that permits operation of commercially-available software products. The data storage management system 100 can be directly connected to a tape library (not shown) and data can be directly transferred between disk and tape.

[0056] In one embodiment, the data storage management system 100 may operate in a different manner for accesses of unprocessed virtual volumes and of virtual volumes implemented with file systems. For unprocessed virtual volumes, backup software typically runs on a server and sends extended block copy commands to the data storage management system 100 to directly transfer blocks between virtual volumes and tape. Since the data that is backed up does not traverse the network to extend to the server and return again from the server to the tape library, server network bandwidth is greatly reduced, and the server is not burdened with the backup task. Volume backup is also facilitated by virtual volume cloning.

[0057] The data storage management system 100 can be managed by a web-based console that allows secure access to system configuration and error status information. The data storage management system 100 may include or use processing capability that determines physical disk and logical disk storage capabilities and saves information relating to storage characteristics in a logical disk characterization structure.

[0058] FIGURE 5 illustrates the creation of a logical disk 106 from the storage space of multiple physical disks 112. The logical disks 106 (e.g., LD4 and LD2) are a first level of

abstraction above physical disks 112 (e.g., PD0, PD1, and PD2). In one embodiment, this can be performed by the volume manager 102 as part of the creation of the virtual volume.

[0059] The volume manager 102 creates logical disks 106, for example by implementing or imposing redundant arrays of independent disk (RAID) groups over the physical disks 112. RAID is a disk redundancy technique using two or more disks that provides continuous data availability and improved performance. Any RAID type may be used including RAID0, RAID1, RAID2, RAID3, RAID4, and/or RAID5. The different RAID types are distinguishable in their methods of data storage and recovery. In one embodiment, the physical disks 112 may be configured as RAID5 in one or more virtual storage regions and configured as RAID1 in one or more other virtual storage regions.

[0060] Physical disks 112 are divided into chunklets 111, which are defined as physical disk portions or segments. In one embodiment, these portions can be of fixed-size, such as, for example, 2 gigabytes (GB) each. In other embodiments, the chunklets 111 may have different sizes (i.e., the chunklets 111 may be a size that is not fixed, or different chunklets 111 may have different sizes). A number of adjacent located chunklets may form a “stripe” 110 (as seen in FIGURE 4).

[0061] Chunklets 111 within a RAID group may be chosen from different physical disks 112 in different disk sleds so that the failure of any physical disk or the removal of any sled does not make data unavailable. Chunklets 111 may also be chosen from physical disks on different link controllers, such as different Fibre Channel controllers, and on different nodes 12 to spread bandwidth across multiple link controllers and nodes 12. Spreading the bandwidth across multiple controllers or nodes results in higher performance and fewer “hot spots,” which are concentrations of high data traffic. The spreading of bandwidth across multiple controllers and nodes is not required for availability in case of link controller or node failure since all physical disks 112 may be connected via a second link loop, to an alternate controller on an alternate node.

[0062] The arrangement of the physical disks 112 into chunklets 111 provides several advantages. One advantage is that configuration into chunklets 111 hides variations in disk size. This allows the data storage management system 100 to use multiple physical disk sizes in the same system without wasting physical disk space. Usage of different physical disk sizes facilitates upgrades using newer disk technology (which typically has greater storage capacity). Another advantage is that RAID configurations can be implemented simply as an arrangement of

chunklets 111. A further advantage is that the configuration allows rearrangement of busy storage areas into chunklets 111. The configuration also permits spacing on a chunklet basis.

[0063] Logical disks 106 are formed by grouping chunklets 111 or stripes 110, for example, into RAID groups. As depicted, a logical disk 106 can be associated with chunklets 111 of different physical disks 112. For example, logical disk LD4 is made up of chunklets PD0:0, PD1:1, PD1:5, and PD2:5 from physical disks PD0, PD1, and PD2. Likewise, the chunklets of a physical disk 112 may be distributed throughout multiple logical disks 106. For example, the chunklets of physical device PD0 are part of or associated with logical devices LD4 and LD2. After creating logical disks 106, the volume manager 102 may create mappings 108 from the regions of the virtual volume 104 (e.g., REGION 0 through REGION N) to one or more logical disks 106 and/or cached block locations in cluster memory.

[0064] Other data access management techniques, such as a virtual table of contents, can also be used. A virtual table of contents may store data and information about the logical disks 106, the physical disks 112, the chunklets 111, and other aspects for implementing the virtual volumes 1026.

[0065] This information may include, for example, the performance characteristics of physical disks 112. In general, the performance characteristics of a physical disk 112 vary according to specific sectors of the disk. An inner disk sector (i.e., one that is closer to the center of the disk) is characterized by lower speeds for transferring data, but shorter search times. In contrast, an outer disk sector (i.e., one that is further from the center of the disk) is characterized by higher speeds for data transfer, but longer search times. Accordingly, the data for any operations that involve large but infrequent transfers of data is ideally stored on outer disk sectors of a physical disk 112, whereas data for any operations that involve smaller but more frequent transfers of data is ideally stored on inner sectors of a physical disk 112.

[0066] A virtual table of contents may save information relating to storage characteristics in a logical disk characterization structure. In one embodiment, for example, the virtual table of contents stores a description of the data arrangement of chunklets 111 for accessing data used for various functions and operations. Each chunklet 111 may be associated with a particular sector of the physical disk 112, and accordingly, may have different storage characteristics. The data storage management system 100 may use the virtual table of contents to match physical and logical storage characteristics to the data requirements.

[0067] In an illustrative example for a storage control method, various characteristics of storage are determined for particular physical disk types and locations (e.g., chunklets 111) and stored in a structure, such as the virtual table of contents. Data handling requirements may also be stored in virtual table of contents. Data for access at high speed is stored in chunklets 111 on the outer disk sectors of a physical disk 112. Other data can be stored in chunklets 111 on the inner disk sectors of a physical disk 112. In one embodiment, this can be accomplished using RAID1 rather than RAID5 protocol. For data that is frequently updated, RAID1 is superior to RAID5 because updates made using RAID5 are more time consuming. For data requiring high reliability, data is stored in physical disks with redundant power supplies.

[0068] As such, a virtual table of contents may be utilized by, for example, a data management system 100, to take advantage of the differences in storage characteristics of physical devices, thereby optimizing or improving overall performance. In typical use, the virtual table of contents is read frequently, but only relatively small amounts of data are transferred. Thus, in an illustrative example, the virtual table of contents may itself be positioned on an inner disk sector of the physical disk 112. Thus, access of the virtual table of contents is improved by permitting more efficient searches for specific data.

[0069] With a virtual table of contents, the virtual volume region tables 104 can be configured to efficiently store information in the storage structures (e.g., logical disks and physical disks) based on characteristics of storage processing. One example of a processing transaction is on-line transaction processing (OLTP). Business organizations maintain mission-critical information in databases for primarily performing on-line transaction processing operations such as order processing. The OLTP databases may be large but typically are not huge in comparison to various enormous data sets that are becoming operational at present. Disk accesses for OLTP are usually characterized by large numbers of small, random operations, typically reads and writes. OLTP operations are almost always considered mission-critical, and often are required to be constantly operational. OLTP is very processor and memory intensive, and database clustering is sometimes used to supply additional computation power. The primary use of clustering is usually to ensure continual availability in the event of a server crash or failure. The system can most effectively configure for OLTP by configuring data storage as chunklets mainly positioned at the inner sectors of multiple physical disks 112 so that disk latencies are low and data searches are fast. Small portions of data are accessed with the OLTP accesses so that high disk access speeds for large amounts of data are not needed.

[0070] FIGURE 6 illustrates the access to the virtual volumes 1026 of multiple nodes 12 by a host device 18 through the virtual volume region tables 104 on several nodes 12. Each host device 18 may have one or more virtual volume region tables 104 for mapping to respective virtual volumes 1026. Each virtual volume region table 104 may be stored locally at its associated node 12 (owner node) and also one or more backup nodes 12 (replicant nodes). The virtual volume region tables 104 provides mappings between the respective virtual volumes and one or more logical or physical disks. Each node 12 may use its virtual volume region tables 104 to update and manage data stored in the physical disks.

[0071] In an illustrative embodiment, a host device 18 may use the virtual volume region tables 104 of any node 12 to which the host device 18 is connected. Thus, for example, if Host 1 is connected to Node 0 and Node 1 (as shown in FIGURE 1), then Host 1 may use table VV_RT (Node 0) or table VV_RT (Node 1).

[0072] In one embodiment, the nodes 12 may use its virtual volume region tables 104 as “hash” tables to perform a hashing operation. That is, a virtual volume region table 104 may implement a hash function, such as a transformation h from an input index m to a fixed-size string $H(m)$. Hash functions can have a variety of general computational uses, and may be used to identify data owner nodes and data replicant nodes, for example in a cache lookup operation. Each node 12 may be designated as an owner node or a replicant node for a set of storage devices (e.g., disk drives). The virtual volume region table 104 may hold an array of control indices or virtual volume offsets that map data to physical storage locations, such as the physical disks 112. Entries in a virtual volume region table 104 may identify nodes 12 that control and store owner and replicant tags that define a location of data storage on physical disks 112 and redundant paths for accessing the data.

Snapshots

[0073] As data and information is stored into the various virtual volumes 1026 in the storage system supported by data management system 100, one or more “snapshots” may be taken of each virtual volume 1026 to record the history of what has been stored in that virtual volume 1026. A snapshot can be a point-in-time picture of the virtual volume at the time that the snapshot is taken. A snapshot can record the state of saved memory including the contents of all memory bytes. Snapshots of the virtual volume 1026 may be used to restore the data storage system in the event of failure. For example, snapshots enable previous versions of files to be brought back for review or to be placed back into use. Snapshots of the virtual volume 1026 can

be taken at regular intervals, or based upon particular triggering events (e.g., upon some indication that the system is about to crash).

[0074] In one embodiment, any data changes in a base virtual volume after an initial point in time may be reflected in a snapshot. Thus, each snapshot may reflect the difference between what is stored in the virtual volume 1026 at one moment of time versus another moment of time. A first snapshot of the virtual volume may correspond to the state of the base virtual volume of data (and mappings) at a time X. A second snapshot of the virtual volume may correspond to the state of the base virtual volume (and mappings) at a time Y. In some cases, any changes or writes of data to the base virtual volume between time X and time Y can be determined by comparing the first snapshot to the second snapshot.

[0075] As described herein, data management and control techniques use data stored as snapshots at various nodes 12 to resolve conflicts, determine precedence of modified data, and provide consistent mirroring of data in virtual volumes 1026 accessed through the nodes 12.

[0076] FIGURE 7 illustrates a number of data structures that may be created and modified as data is written or changed in a base virtual volume 654. As depicted, these data structures include a number of tables arranged in a hierarchy of multiple levels (e.g., Level 1, Level 2, and Level 3). At Level 1, there is a table 666a. At Level 2, there are tables 668a and 668b. At Level 3, there are tables 670a, 670b, and 670c. Although three levels are shown, it should be understood that in other embodiments the hierarchy may comprise any number of levels. The base virtual volume 654 may be the most current version or state of a virtual volume 1026.

[0077] The tables 666, 668, and 670 may be used to track any data changed or written to the base virtual volume 654 for one or more snapshots. As depicted, four write operations are made to write data (e.g., W1, W2, W3, and W4) into various parts of the base virtual volume 654. In the illustrative example, each of these write operations to the base virtual volume 654 cause the data which was modified to be captured in a snapshot, which can be the same or different snapshots for the various write operations. Data for the various snapshots is stored in data storage areas 602, 604, 606, 608, which can be in physical storage devices (e.g., physical disks 112) or in virtual volume memory. Each data storage area 602, 604, 606, or 608 can be a page of data which, in one embodiment, may comprise one or more disk or data blocks. A data block can be the minimum size or region of data accessible from a physical storage device. Taken together, the tables at the various levels and the data storage areas may provide a snapshot of data written to multiple physical disks using virtualized disk operations. In some embodiments,

creation and modification of the tables 666, 668, and 670 may be controlled by a “master” node 12, which has a backup master node in case the master node malfunctions or is inaccessible.

[0078] In an illustrative embodiment, Level 1 (or L1) table 666a is a first level mapping structure. The table 666a comprises a plurality of entries (e.g., 1024). Each entry in the Level 1 table 666a addresses a range of memory locations (or segments) of the entire base virtual volume 654. Each segment may, for example, comprise 1000 memory locations of the entire base virtual volume 654. Assuming that there are 10 entries in the Level 1 table 666a, a first entry (entry 0) addresses locations 0-99 of the base virtual volume 654 (which may correspond to a first Level 2 table), a second entry (entry1) addresses locations 100-199 (which may correspond to a second Level 2 table), and so on. Each Level 2 table 668a, 668b may comprise a plurality of entries (e.g., 10), each of which corresponds to a particular range of memory locations within the segment of the Level 1 entry pointing to that Level 2 table. For example, a first entry of a Level 2 table may address locations 0-9 of the base virtual volume 654 (which may correspond to a first Level 3 table), a second entry of the same table may address locations 10-19 (which may correspond to a second Level 3 table), and so on. Each Level 3 table may also comprise a number of entries (e.g., 10), each of which points to a particular storage area (e.g., 602, 604, 606, or 608) storing data that was changed or written. In one embodiment, each Level 2 table is controlled by a specific node 12, which may also control the Level 3 tables associated with that Level 2 table.

[0079] The structures and techniques described herein are highly suitable for identifying storage locations and accessing widely separated and sparsely concentrated physical storage devices accessible by a virtual volume. Data snapshots typically involve changes to only a small portion (e.g., 1% or less) of the entire storage space of virtual volume, where the data changes occur at locations that are generally widely separated. In one embodiment, data structures for the snapshots are recursive so that further tables for snapshot volume are created only when write accesses are made to those particular levels.

[0080] This can be accomplished by the volume manager accessing and attempting to check entries in the tables 666, 668, and 670 to determine whether a particular physical block has previously been written. For example, in an illustrative embodiment, if data is written to a storage location of the base virtual volume 654 that falls within a particular segment, the volume manager 102 first reads the entry of the Level 1 table 666 that corresponds to the target physical storage. If the entry is null/empty, thus indicating no previous writes to that location, the volume manager 102 sets a pointer in the corresponding entry of the Level 1 table 666a and creates a

Level 2 table 668. The pointer in the Level 1 table 666a points to an entry in the Level 2 table (e.g., table 668a or 668b) for further specifying the specific location of the base virtual memory 654. Tables and corresponding entries for other levels (e.g., Level 3 table 670) are generated and made. An entry in the final level table specifies the storage location, which may be in a virtual volume 1028. With the entries, pointers are set in the various tables, and a page of data is written to the physical storage area. Alternatively, if the entry of the Level 1 table 666 is not null/empty, the volume manager 102 designates or creates a pointer to an existing element of a Level 2 table (e.g., table 668a or 668b). The volume manager 102 reads the element of the Level 2 table 668 that corresponds to the target physical storage. If this entry is null/empty, the volume manager 102 creates a new Level 3 table 670; otherwise, the volume manager 102 uses the element as a pointer to an existing Level 3 table. This is continued for all level until a page of data is written to the physical storage area.

[0081] The various level mapping tables (e.g., Level 1, Level 2, and Level 3 tables) may be considered exception tables. This is because, in some embodiments, entries in the level mapping tables only exist if data has been change or written (which is considered an exception, rather than the norm) in the respective storage areas.

[0082] The state of data for the virtual volume 1026 at a specific time can be brought back or placed into use by accessing the data for a snapshot created at that time.

[0083] In one embodiment, a copy-on-write (COW) technique can be used in conjunction with the snapshots. In a COW technique, a data page or block is copied to a snapshot before that data block is modified by any write operations. Generally, only the first write operation to a given data block causes a COW operation ("a COW push") to a snapshot. Subsequent write operations to that data block are allowed to directly change the base virtual volume 654. Alternatively, a complete copy of all the data blocks is made to the snapshot. After the complete copy, all of the data blocks can be modified.

[0084] The tables 666a, 668a, 668b, 670a, 670b, and 670c can illustrate an example of copy-on-write operations that result in data (W1, W2, W3, and W4) being stored in the data storage areas 602, 604, 606, and 608. A first data write operation to a location in the base virtual volume 654 causes a copy of the data W1 to be stored in storage area 604. Storage area 604 is addressed by the tables 666a, 668a, and 670a. A second data write operation causes a copy of data W2 to be stored in storage area 606. The second data storage area 606 is addressed by some of the same tables as storage area 604 (i.e., tables 666a and 668a) and a different table

(table 670b). A third data write operation causes a copy of data W3 to be stored in storage area 608. The data storage area 608 is addressed by tables 666a, 668b, and 670c. A fourth data write operation causes a copy of data W4 to be stored in storage area 602. The data storage area 602 is addressed by the same tables as storage area 604 (i.e., 666a, 668a, and 670a).

[0085] The copy-on-write technique can be implemented by creating and managing the hierarchy of tables 666, 668, and 670 which ultimately point to data storage areas 602, 604, 606, 608. Specifically, for each copy-on write operation, a volume manager 102 may determine if a particular area of storage was previously written. If the area was not previously written, the volume manager 102 creates appropriate tables at the various levels for that storage area. Otherwise, if the area was previously written, all appropriate tables should already have been created, and the volume manager 102 functions to add or modify entries in the tables to reflect the data changes to the virtual volume 654.

[0086] In one embodiment, snapshot techniques can be used in conjunction with cloning techniques in the data storage system. The data storage management system 100 may generate remote mirror copies or “clones” of data on virtual volumes 1026 and logical disks 106 in the multiple nodes 12. The system 100 manages remote mirror cloning of data segments of a virtual volume 1026 by creating local and remote mirror data structures (which may include various level tables and snapshot data). When a clone is first created, the system allocates storage space for a clone structure resident on or controlled by a remote node 12 that corresponds to the data structure in the local node 12. The system stores header or management information that defines the local and remote structures as mirror copies. When data is subsequently written to the storage structures of one of the local or remote nodes 12, information is transferred to the other node 12 so that the same data is written to the clone.

[0087] In one embodiment, the data structures (e.g., L1, L2, L3 tables and W1, W2 data storage spaces or pages) for snapshots of a virtual volume 1026 can be stored various volumes of memory (which themselves can be virtual or real).

[0088] FIGURE 8A illustrates one view of storage volumes for the data structures for a snapshot technique. As depicted, these may include a base virtual volume 654, snapshot administration volume 656, and snapshot data volume 658 for various snapshots.

[0089] The base virtual volume 654 may be the most current version of a virtual volume 1026. Thus, the base virtual volume 654 may comprises data stored in the virtual volume at

some initial point in time, such as time X (or time 0), and any data that has been subsequently written by a host after time X. The base virtual volume 654 is associated with a virtual volume region table (e.g., virtual volume region table 104) that maps regions of the base virtual volume 654 to physical storage devices (e.g., physical disks 112). The base virtual volume 654 may specify or include virtual data structures of all physical storage devices in communication with a plurality of nodes 12 in a data handling system 10. As an example, a base virtual volume 654 of a multi-node system may comprise 1 Terabyte (“1T”) of data. As data in the base virtual volume 654 is changed or modified over time, various snapshots can be taken to provide a history of what has been stored in that virtual volume at different moments.

[0090] Snapshot data volume 658 stores data for each snapshot--i.e., data that has been written/changed in the base virtual volume 654 from an initial point in time to when a snapshot is taken. As depicted, separate data may be stored for each snapshot of the virtual volume. The snapshot administration volume 656 stores a number of tables 666, 668, and 670 in a hierarchy with multiple levels (e.g., Level 1, Level 2, and Level 3). The different levels of tables may map the data of a snapshot back to a particular location of the virtual volume (as described with reference to FIGURE 7), so that the state of the base virtual volume at previous point in time can be re-created.

[0091] FIGURE 8B illustrates another view of storage volumes for the data structures for a snapshot technique. Similar to the view depicted in FIGURE 8A, these volumes include a base virtual volume 654, a snapshot administration volume 656, and snapshot data volume 658. Snapshot data (reflecting data changes made to base virtual volume 654) may be stored in any space which is available and accessible in snapshot data volume 658. Likewise, snapshot tables (which map the snapshot data back to the base virtual volume 645 for particular snapshots) may be stored in any available and accessible space of snapshot administration volume 656.

[0092] The tables for multiple levels (e.g., Level 1, Level 2, and Level 3) may each contain entries for various snapshots. As depicted, an entry 680 of the Level 1 (or L1) table comprises a virtual volume name (VV_name) and an offset. This virtual volume name is for the Level 2 (or L2) table and the offset points to a particular entry 682 in the table. This entry 682 for the L2 table also comprises a virtual volume name and offset. The volume name identifies the Level 3 (or L3) table and the offset points to a specific entry 684. The L3 entry 684 comprises a virtual volume name for the snapshot data volume 658 and an offset which points to specific data (e.g., data page).

Base and Snapshot Volumes

[0093] In one embodiment, a tree-like data structure formed from or comprising a base virtual volume and a series of snapshot volumes stores data and tables for various snapshots of a virtual volume. Each snapshot volume maintains data and tables for an associated snapshot of the base virtual volume. A snapshot volume can be either a read only snapshot volume (“ROSS”) or a read/write snapshot volume (“RWSS”). A ROSS presents a constant view of the data in a virtual volume at a specific time. After creation of a ROSS, data can be read from but not written into the ROSS. A RWSS descends from a ROSS (e.g., a parent snapshot volume) and may serve to hold modifications to the parent ROSS. A RWSS can be read and written like a base virtual volume. A snapshot volume may be created or started by execution of a command from the volume manager 102, a node 12, or a host device 18. All data writes are collected in a particular snapshot volume until a subsequent snapshot volume is created or started. The base virtual volume can be accessed to obtain the newest or most current data. Any of the snapshot volumes can be accessed to obtain data that was written at a prior time.

[0094] In one embodiment, in response to a write to a location in a base virtual volume, a copy-on-write (COW) operation may cause creation of a first snapshot volume. Subsequent writes to the base virtual volume or to the first snapshot volume can similarly result in creation of additional snapshot volumes over time. For example, a second snapshot volume stores data that has been more recently written than data in the first snapshot volume. If return to a particular state of memory is desired, a snapshot volume corresponding to the snapshot for the particular state is accessed. In this way, copy-on-write operations can be reversed.

[0095] With the tree-like data structure, a differential backup operation can be readily performed. In particular, the differences between any two snapshots in a branch can be determined to identify or locate modified pages. Furthermore, it is easy to tell how a page is shared among snapshot volumes and a base volume. This makes a cache lookup more efficient.

[0096] Additional details regarding the tree-like data structure and its advantages are provided in co-pending U.S. Provisional Patent Application No. _____, entitled “Read-Write Snapshots,” (Attorney Docket No. 3PD-P100) filed concurrently. Such application is assigned to the same Assignee as the present application and is hereby incorporated by reference in its entirety.

[0097] In an illustrative embodiment in a scalable data handling system, the owner node of a target region of the virtual volume determines whether there is an entry for the region in the

most recent snapshot volume. If an exception list (i.e., pointer tables) of the most recent snapshot volume does not have an entry for the target block, then the owner node reads the data at target virtual volume region in the base virtual volume, creates appropriate tables and entries in the exception list of the most recent snapshot volume, and writes the old base volume data to the snapshot volume. The owner node writes the new data block to the base virtual volume and sends an acknowledge signal to the node that received the write request.

[0100] In another illustrative embodiment, the host device 18 or system writes data by indexing into the virtual volume region table 104 to determine which node 12 is the owning node and which node 12 is the replicant node for the intended data. The write operation is executed and a copy of the data is sent to the owning and replicant nodes. The host device 18 or system uses the virtual volume region table 104 to determine the location of the intended replicant so that, in case a block write operation fails, the host device 18 can access the redundant copy of the data.

[0101] The owner node is responsible for maintaining coherency of data. In one embodiment, the owner node maintains coherency by managing the control and data structures (described herein) that specify the location of data blocks, and the virtual volume region table 104 that specifies the nodes 12 responsible for particular data blocks. The owner node sends messages to other nodes, informs the other nodes of access to the owner node's physical storage, and requests updating of tables at the other nodes. The owner node waits for the other nodes to acknowledge the message to ensure that all nodes have coherent tables. Thus, consistent tables are maintained across multiple nodes.

[0102] In one embodiment, if a node 12 fails, the volume manager 102 uses the virtual volume region table 104 to provide access to redundant data of the virtual volume through the replicant node. The replicant node may access the virtual volume region table 104 and other data management structures to determine how to derive information in case of failure. For example, the replicant node can access information such as transaction logs for error recovery when a failure requires the replicant node to assume management conditions of a failing owner node.

[0103] In one embodiment, data of a virtual volume may be read in a manner opposite to a write, for example, by accessing the data from the base virtual volume and one or more snapshot volumes, as desired. A component of the data storage system reads a data block from the base virtual volume by simply reading the physical storage designated by the base virtual volume

mappings. To read a data block from a snapshot volume, the system first examines the oldest snapshot volume to determine whether the requested data block was written or modified in the oldest snapshot volume. The system makes the determination by accessing the exception list of the oldest snapshot volume. A snapshot volume is read by reading through the level lists from the Level 1 table 666 to the last level table (e.g., Level 3 tables 670a, 670b, or 670c). If the pointers in the tables of the oldest snapshot volume associated with the data block are not zero, then an entry for that element exists on the physical volume, and the pointers are used to read the stored data from that snapshot volume.

[0104] If a pointer in any table of the oldest snapshot volume for the requested data block is zero, data has not been written to the read target block. For example, if the target entry in a level list is zero, then no data corresponding to the target element exists. Then the system examines the next oldest snapshot volume to see whether the data block was modified in during that snapshot, and so on until it reaches the most recent snapshot volume. The snapshot volumes are read until a pointer value that is not zero is available or the base virtual volume is reached.

[0105] In one embodiment, data of the base and snapshot volumes may be exported or transferred between nodes 12 of the data storage system. The system passes level by level through the storage structures until a pointer is available, and reads the data designated by the first available pointer. If the data block was not found in any of the snapshot volumes, then the system looks in the base virtual volume.

[0106] In one aspect, if a host device 18 issues a read request to an owning node 12, the owning node 12 responds by sending the requested data directly to the host device 18 without data shifting. By avoiding data shifting, the read operation, as well as a corresponding write to an owning node 12, can be 35 to 40% faster.

[0107] In some embodiments, pointers may be set to skip over one or more snapshot volumes. For example, if a desired data block is found in the fourth snapshot volume, then a pointer may be set in the first snapshot volume so that a subsequent search for the data block in the first snapshot volume will automatically skip to the fourth snapshot volume. This saves time and improves performance by avoiding the second and third snapshot volumes in subsequent searches for that data block.

[0108] In another embodiment, a snapshot read operation is performed by first accessing the data structures of the most recent snapshot volume before the data structures of the base virtual

volume so that the latest written data is accessed. In a read of the snapshot volumes, the system searches the various Level 1, 2, and so on tables (FIGURE 7), and if a pointer entry is found in the snapshot volume, the entry is returned as the result of the read operation. A pointer in the final level table (e.g., Level 3 tables 670a, 670b, or 670c) points to a block in physical storage. If no pointer entry is found in the snapshot volumes, the system returns to the base virtual volume.

Exemplary Methods

[0109] FIGURE 9 is a flowchart for an exemplary method 800 for a multiple level mapping for a virtual volume 1026, according to an embodiment of the present invention. In one embodiment, method 800 may be performed by volume manager 102 (FIGURES 3 and 4). This method may cause tables at various levels (e.g., Level 1, Level 2, and Level 3) to be generated or created. Method 800 begins at step 802 where volume manager 102 allocates a Level 1 mapping table for the virtual volume.

[0110] At step 804, a write operation is initiated. This write operation may be directed to a particular storage segment or location in the virtual volume. The volume manager 102, at step 806, looks up for an entry in the Level 1 mapping table corresponding to the segment or location. At step 808 the volume manager 102 determines if an appropriate entry for the segment or location exists in the Level 1 mapping table.

[0111] If no entry exists, then this location of the virtual volume has not been written to previously for a present snapshot, and accordingly, no Level 2, Level 3, etc. mapping tables would have yet been created or allocated. At steps 812 and 814 a suitable Level 2 mapping table is allocated and an appropriate entry is created. Then at steps 820 and 822 a suitable Level 3 mapping table is allocated and an appropriate entry is created. At step 828 a copy-on write (COW) page for the data is created, after which method 800 moves to step 830.

[0112] Alternatively, if at step 808 an appropriate entry is found in the Level 1 mapping table, then the volume manager 102 accesses the Level 2 mapping table to which the Level 1 entry points. At step 810 the volume manager 102 looks for an entry in the Level 2 mapping table corresponding to the particular segment or location of the virtual volume. At step 816 the volume manager 102 determines if an appropriate entry for the segment or location exists in the Level 2 mapping table.

[0113] If no entry exists in the Level 2 mapping table, then method 800 moves to steps 820 and 822 where a suitable Level 3 mapping table is allocated and an appropriate entry is created. Thereafter method 800 moves to step 828. Otherwise, if at step 816 an appropriate entry is found in the Level 2 mapping table, then the volume manager 102 accesses the Level 3 mapping table to which the Level 2 entry points. At step 818 the volume manager 102 looks for an entry in the Level 3 mapping table corresponding to the particular segment or location of the virtual volume. At step 824 the volume manager 102 determines if an appropriate entry for the segment or location exists in the Level 3 mapping table.

[0114] If no entry exists in the Level 3 mapping table, then method 800 moves to step 828, where a COW page for the data is created. Otherwise, if at step 824 an appropriate entry is found in the Level 3 mapping table, then the volume manager 102 obtains the COW page to which the Level 3 entry points.

[0115] At step 830, the COW page is updated. Thereafter, method 800 ends.

[0116] Accordingly, the various mapping tables (for Level 1, Level 2, and Level 3) provide the management of COW pages for the virtual volume. Because at least some of the mapping tables are not allocated until they are needed, disk resources are only used or committed when a COW happens and memory resources are committed only when a reference to a particular storage area is made.

[0117] FIGURE 10 is a flowchart for an exemplary method 900 for a delayed copy-on write (COW) operation for a virtual volume, according to an embodiment of the present invention. In one embodiment, method 900 may be performed by a volume manager 102. Method 900 begins at step 902 where a host device (e.g., a server computer) initiates a write to the virtual volume (i.e., initiates a write operation) which triggers a copy-on write (COW) operation.

[0118] A COW operation may involve a read operation and possibly a merge operation. A read from disk storage can take a relatively long time; a merge may be faster. In previously developed systems, the read and merge operations may delay a host device by causing an interrupt or “CPU busy-copy.” As a result, if the write operation of a particular host device triggers a COW, the response time to the host device could be relatively long because a COW must first be performed.

[0119] The method 900 addresses this problem as follows. At step 904, an acknowledgement of the write is made to the host device. This allows the host device to

continue. At step 906 the copy-on write operation is completed, which may involve, for example, looking up entries in various level mapping tables (e.g., Level 2 or Level 3 table), the allocation of one or more level mapping tables and creation of entries for the same, or the creation and updating of a COW page. During completion of the COW operation, the volume manager may signal or indicate “COW pending.” Since the host device is not waiting for the COW operation to be completed, overall system latency is reduced. Thereafter, method 900 ends.

[0120] With the delayed COW technique, according to one embodiment of the invention, even if a write operation of host device triggers a COW operation, the write operation is acknowledged to the host device without the actual completion of a COW operation. Thus, processing or performing the COW operation is invisible to the host device, which provides for better response time.

[0121] One advantage of the usage of virtual volumes and snapshot volumes as described is that a large data capacity can be provided in a data storage system. An arbitrarily large number of data blocks can be stored. Another advantage of the virtual volume and snapshot volume data storage and handling technique is a very small impact on performance in the form of access time. Yet another advantage of the storage and handling techniques is that performance is consistent for all variations of data density so that sparsely as well as densely distributed entries all have relatively similar access times. Consistent performance is more predictable and stable in a multi-tasking, multiple-thread, or multi-use environment for the data storage system. A further advantage is that with the virtual volume and snapshot volume data structure, data backups are efficiently and easily made by storing only differences in data, e.g., exceptions from a base virtual volume.

[0122] While the invention has been described with reference to various embodiments, it will be understood that these embodiments are illustrative and that the scope of the invention is not limited to them. Many variations, modifications, additions and improvements of the embodiments described are possible. For example, those skilled in the art will readily implement the steps necessary to provide the structures and methods disclosed herein, and will understand that the process parameters, materials, and dimensions are given by way of example only and can be varied to achieve the desired structure as well as modifications which are within the scope of the invention. Variations and modifications of the embodiments disclosed herein may be made based on the description set forth herein, without departing from the scope and spirit of the invention as set forth in the following claims.